**The Verification and Scoring of Weather Forecasts**

Iriving I. Gringorten

*Journal of the American Statistical Association*, Vol. 46, No. 255 (Sep., 1951), 279-296.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28195109%2946%3A255%3C279%3ATVASOW%3E2.0.CO%3B2-U

*Journal of the American Statistical Association* is currently published by American Statistical Association.

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

*Number 255* SEPTEMBER 1951 *Volume 46*

# THE VERIFICATION AND SCORING OF WEATHER FORECASTS

IRVING I. GRINGORTEN
*Air Force Cambridge Research Center*

In scoring a forecast, one of two purposes should be considered: either to determine the utility of the forecast or to determine the skill of the forecaster. If the purpose is utility, then the score for each forecast should be directly proportional to the value of the forecast in meeting specified operational requirements. If skill is the purpose, then the total score for a series of forecasts should be a reflection of the forecaster's ability to analyze and classify a weather situation for forecasting purposes; within a well-defined class of antecedent weather the probability of a subsequent event, like rain, is increased above (or decreased below) the relative frequency of that event in the total of all weather situations, as established in a long period of time.

## I. INTRODUCTION

MUCH has been written on the verification of forecasts. While some have minimized the importance of adhering to a system of verification, most authors have recognized its necessity. With no system of scoring, the usefulness of forecasting, generally speaking, would remain in doubt; the forecasters themselves would not know if, or when, they show improvement in their art, or whether one forecaster is doing better than another. Two questions are asked repeatedly: Is it better to perpetually forecast the most frequent event, say, "no rain," if it will average better than a supposedly skillful prediction? Will the forecaster be correct most frequently if he forecasts persistence, that is, today's weather repeated as the forecast for tomorrow? As long as these questions are open, to the embarrassment of the forecaster, then the last report on verification and scoring has not been written.

There have been several surveys of the literature [1, 2, 3] reviewing some 55 papers on verification. The earliest papers were written in 1884

[4, 5, 6, 7]. More recently, Glenn Brier [8, 9] has devised systems to verify probability forecasts. An excellent critique has been written by personnel of the USAF Air Weather Service [10], but not published. While it is not my intention to review the previous papers, it seems advisable to cite the key points:

1. It is desirable to determine the utility of a forecast and to determine at what point it ceases to have any value [7, 11].

2. Consideration should be given to predictions of non-occurrence as well as predictions of occurrence [5, 7].

3. Forecasts should be compared with "blind" forecasts to determine whether the forecasts are better than pure guesswork or better than those forecasts obtained by a non-skilled rule of thumb. From the standpoint of skill, percentage of "hits" *per se* is meaningless [6, 12].

4. The weight (or score) attached to a forecast should be affected by the importance of the forecast, the difficulty of making an accurate forecast, and the proximity of the forecast to the verification [7, 10, 11, 12, 13, 14, 15].

5. For a suitable system of verification and scoring, forecasts should be clearly and unambiguously stated so that they can be checked quickly and accurately against the subsequent events [general agreement among the authors].

6. A forecast will have greater value if the forecaster becomes acquainted with the specific operational requirements [16].

7. Skill in forecasting has been defined as the number of hits in excess of those obtainable by chance, or by forecasting persistence, or by perpetually forecasting the event that has occurred most frequently in the past [12].

The above principles are acceptable for this paper except for two key points. First, *a clear distinction should have been made between the utility of a forecast and the skill of the forecaster.* Failure to emphasize this distinction has undoubtedly caused much of the difficulty of verification and scoring. Secondly, skill should be redefined, as is done in the second paragraph below.

The *utility* of a forecast should be judged by the operational requirement; in fact, the forecast becomes the working assumption. It might well be that the forecaster cannot contribute more than the climatic data and frequencies for given purposes of operation; or the persistence forecast (tomorrow's weather will be the same as today's) might be the best working assumption for certain needs. It is important, therefore, to devise a scoring system so that the score allotted to a forecaster would be an index of the value of his forecast for given operational needs. The prediction of fog at a given airport might carry with it five times as much weight, taken operation-wise, as the predic-

tion of clear skies. Moreover, there might be some weight attached to a "nearly correct" forecast in proportion to its operational value; the prediction of low clouds at the given airport, where fog ultimately develops, might carry with it a definite score because it would have alerted the operations office to the possibility of fog. Lastly, a forecast might become important in the light of the present state of the weather [12]; if the airport is closed by fog now, it becomes important to forecast its time of opening.

In defining skill, this paper departs from previous reports: *Skill is the forecaster's ability to analyze and classify the antecedent weather so that, within one class, the probability of a subsequent event is increased above, or decreased below, the relative frequency of that event in all weather situations* (hereafter referred to as the "climatic frequency"). For example, if the climatic frequency of rain, obtained from 10 years of record, is 5% and the forecaster recognizes in the present situation a 20% probability for rain, then he is exhibiting his knowledge and skill. On the other hand, if we accept as skill "the number of hits in excess of those obtainable through chance forecasts," we would be forced to conclude that, with respect to a rare event of 5% frequency, the forecaster would exhibit no skill at all if he could not obtain more than 95% of his forecasts correct.

A single forecast cannot be made to meet the operational needs and a test of skill at one and the same time, except by pure coincidence. The forecaster should be asked to make two separate forecasts, one for use, and the other for a test of his skill; and the two forecasts may be in disagreement. The operational requirements themselves might call for several forecast statements or working assumptions, aside from any test of skill [3]. The methods of verification and scoring, described below, form two groups, those that test the operational value of the forecasts, and those that test the skill of the forecaster.

## II. METHODS

To proceed further it is convenient to introduce symbols for the events which are to be forecast. Let $X_0, X_1, \cdots, X_n$ denote $(n+1)$ mutually exclusive events. For example, $X_0$ might represent ceiling (i.e. cloud height) zero, $X_1$ might represent ceiling 100 to 400 feet, $X_2$ might represent ceiling 500 to 900 feet, and so on. The symbol $X_k$ denotes any class of event of the kind $X$.

The required form of the forecast might be one of the following: (1) For his forecast, the forecaster must select a single event from the $(n+1)$ mutually exclusive events, but he obtains partial credit for his choice depending upon the nearness of his choice to the observed event; (2) The forecaster can select more than one event as alternate possibil-

ities; but his credit, upon verification, is greater if he has selected fewer alternates; (3) The forecaster must group the $(n+1)$ events dichotomously and select one group, as his forecast. Each of these forms of the forecast is described below:

1*a*. *Requirement: a single choice from* $(n+1)$ *mutually exclusive events; test of operational value.*

The forecaster usually faces a situation in which he sees the chances for several alternate developments. For example, clouds might develop low enough to close the airport, or low enough to restrict the number of flights in and out of the airport, or high enough to permit normal schedules. One of these possible developments must be chosen as the working assumption of the operations office. While the best assumption is the event that verifies, yet there is generally value in an assumption that proves nearly correct.

Let $\alpha_{ki}$ be the score obtained by the forecaster if his prediction was $X_k$ but the subsequence is $X_i$; then $\alpha_{ii}$ is the perfect score; a table of scores would resemble Table 1. For operational purposes the score should be made directly proportional to the net profits resulting

TABLE 1. Scores for each combination of forecast and observed event; antecedent condition: $Xm'$ (schematic)

| Forecast | Observed | | | |
|---|---|---|---|---|
| | $X_0$ | $X_1$ | $X_2$ | $X_3$ |
| $X_0$ | $\alpha_{00}$ | $\alpha_{01}$ | $\alpha_{02}$ | $\alpha_{03}$ |
| $X_1$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ | $\alpha_{13}$ |
| $X_2$ | $\alpha_{20}$ | $\alpha_{21}$ | $\alpha_{22}$ | $\alpha_{23}$ |
| $X_3$ | $\alpha_{30}$ | $\alpha_{31}$ | $\alpha_{32}$ | $\alpha_{33}$ |

from the assumption (or forecast). Sometimes such a score would be negative, representing losses incurred by the operator. The score for an incorrect forecast must not be greater than the score for the correct forecast, so that $\alpha_{ki} \leqq \alpha_{ii}$. It is the job of the operations office to choose reasonable values for the $(n+1)^2$ scores $\alpha_{ki}$. How this might be done is illustrated by the following:

Let us suppose that the rate at which airplaines can be accepted into an airport,[1] like LaGuardia Airport, is governed by the flying weather classification at the airport. That is, for each class $X_k$ of the flying weather (determined by cloud base, cloud top and visibility)

[1] The term "acceptance rate" has been used in this connection by members of the Air Navigation Development Board, a technical advisory group under the Secretaries of Defense and Commerce. The writer is indebted to this group for a description of this problem.

there is a fraction $f_k$ of the maximum rate at which airplanes could enter the airport. Then the forecaster of an airline, essentially, will be forecasting $f_k$, and the operations office simply will want $f_k$ as a working assumption. Let $T$ be the net profit resulting to an airline from each successful flight. Let $L$ be the loss resulting from an incorrect decision or an unsuccessful flight; the loss would be due to overhead, the salaries of the crew, cost of gasoline and loss of good-will expressed in dollars and cents. Let $l$ be the net cost resulting from cancellation; this should be the overhead such as the cost of rentals, maintenance of idle equipment and salaries. Let $\alpha_{ki}$, the score, be identical with the mean net profit resulting from the operation of the fraction $f_k$ of the normal number of flights and the cancellation of the remaining flights. Then

$$(1) \qquad \alpha_{ki} = \begin{cases} f_k T - (1 - f_k)l & \text{for } f_k \leqq f_i \\ f_i T - (f_k - f_i)L - (1 - f_k)l & \text{for } f_k > f_i. \end{cases}$$

If $T=3$, $L=2$, $l=1$ (purely hypothetical), and if there are three acceptance rates: $f_0=0$, $f_1=1/2$, $f_2=1$, then from equations (1) the scores would be those in Table 2.

TABLE 2. Scores that are directly proportional to net profits or losses, if $T=3$, $L=2$, $l=1$, (see text)

| Forecast, or Working, Assumption | Verified | | |
|---|---|---|---|
| | $f_0$ | $f_1$ | $f_2$ |
| $f_0$ | −1 | −1 | −1 |
| $f_1$ | −3/2 | 1 | 1 |
| $f_2$ | −2 | 1/2 | 3 |

Is the above example a realistic one? From the point of view of an airline operator it is undoubtedly oversimplified. But, if a forecaster is asked to predict a single event, then, essentially, he is being asked to make the operational decision, because he cannot predict the weather precisely. If it is admitted that the forecast is, in fact, the working assumption, then it must also be admitted that the forecast must be governed by the economics of the operation. Either the decisions must remain forever subjective, or a table of scores, such as Table 2 should be computed to guide the forecaster or operator in his decisions.

In $s$ trials the forecaster's rating, $F$, is the ratio of his accumulated score to the accumulated score of perfect forecasting:

$$F = \frac{\sum \alpha_{ki}}{\sum \alpha_{ii}}.$$

The forecasting system that maximizes this rating is the preferred system for the given operational requirements. This rating has the maximum value of unity, but may assume negative values since the scores reflect the profits, and oftentimes there are losses instead of profits. If the persistence forecast yields the greatest rating, $F$, then it is proper and fitting to forecast by persistence for the given operations and neglect the skill of the forecaster.

If this author is permitted to digress from the main theme of this paper, it is desirable to show how the forecaster might decide which working assumption is best for the operations office. Faced with his own classification of antecedent weather, the forecaster might realize that, for today's class of weather, each acceptance rate has a certain probability of verification. Within the forecaster's designated class, let $p_0, p_1, \cdots, p_m, \cdots, p_n$ be the true probability of each acceptance rate $f_0, f_1, \cdots, f_n$. (In this discussion the $p_m$'s are not subjective probabilities, although they may be so difficult to determine that, in practice, the forecaster will have to resort to approximations. In theory, each $p_m$ must be assumed to exist as the limit of the relative frequency of the subsequence $X_m$ within today's class of antecedent weather.) Then

$$\sum_{m=0}^{n} p_m = 1.$$

Within the class of weather as analyzed by the forecaster, the expected average profit netted to the management under the working assumption $f_k$ will be proportional to $A_k$ in the equation

$$(2) \qquad p_0\alpha_{k0} + p_1\alpha_{k1} + \cdots + p_k\alpha_{kk} + \cdots + p_n\alpha_{kn} = A_k.$$

Equation (2) represents $(n+1)$ equations, one for each working assumption. The best selection by the forecaster or operations office is that working assumption which maximizes $A_k$.

1*b*. *Requirement: A single choice from (n+1) mutually exclusive events; test of skill.*

To test the forecaster's skill, the method of scoring should force him to select, as his forecast, an event which, in view of the antecedent weather, has a probability of occurrence greater than the climatic frequency of that event. (It could be argued that this is not a forecast in the usual sense of the word, but, undeniably, it is a forecast in the general sense [17].) The following paragraphs are devoted to finding values for the $\alpha_{ki}$'s (Table 1).

Let $P_{c0}, p_{c1}, \cdots, p_{cn}$ be the climatic frequencies of each future event

$X_0, X_1, \cdots, X_n$ for the given period of the year and for the given time of day, as established over a long period of time. Then

$$\sum_{m=0}^{n} p_{cm} = 1.$$

If a forecaster blindly selects $X_k$ as his forecast throughout an arbitrary set of $N$ days, then, for the fraction $p_{c0}$ of the $N$ days, his expected average score would be $p_{c0}\,\alpha_{k0}$; for the fraction $p_{c1}$ his expected average score would be $p_{c1}\alpha_{k1}$, and so on; or, for all $N$ days, his expected average score would be

$$\sum_{m=0}^{n} p_{cm}\alpha_{km}.$$

In accordance with this paper's definition of skill, it is necessary to make it no more profitable for the "blind" forecaster to predict any one event, throughout an arbitrary set of $N$ days, than to predict any other event, that is, to make his expected average score equal to a low constant, say 1.0. Thus,

$$\sum_{m=0}^{n} p_{cm}\alpha_{im} = 1 \qquad \text{for any } X_i (0 \leqq i \leqq n). \tag{3}$$

There are several relations that are self-evident. First, the score for a partially correct forecast should be always less than the score for the totally correct forecast. Or,

$$\alpha_{ki} < \alpha_{ii}. \tag{4}$$

Secondly, the score for a completely erroneous forecast should be zero. Or,

$$\alpha_{0n} = \alpha_{n0} = 0. \tag{5}$$

Let us suppose that a forecaster had predicted consistently *one* future event for a class of $N$ days, but the true probabilities of $X_0$, $X_1, \cdots, X_n$ within that class of $N$ days are $p_0, p_1, \cdots, p_n$. To begin with, we may suppose that the true probability of $X_k$ increased at the expense of the others. That is,

$$p_k > p_{ck} \quad \text{and} \quad p_m \leqq p_{cm} \qquad \text{for } m \neq k.$$

But since

$$\sum_{m=0}^{n} p_m = \sum_{m=0}^{n} p_{cm} = 1$$

$$p_k - p_{ck} = \sum_{m \neq k} (p_{cm} - p_m) \geqq (p_{cm} - p_m) \qquad \text{for } m \neq k. \tag{6}$$

By the definition of skill, and accepting the value of 1.0 for no skill, it is necessary that the forecasting of $X_k$ on each of the $N$ days, when $N$ is large, should net the forecaster a true average score greater than 1.0, and the forecasting of anything else should net him a true average score of 1.0 or less.
That is,

$$\sum_{m=0}^{n} p_m \alpha_{km} > 1 \tag{7}$$

and

$$\sum_{m=0}^{n} p_m \alpha_{im} \leqq 1 \qquad i \neq k. \tag{8}$$

From (7) and (3)

$$(p_k - p_{ck})\alpha_{kk} > \sum_{m=0}^{n} (p_{cm} - p_m)\alpha_{km} \qquad \text{for } m \neq k. \tag{9}$$

In view of relation (6), relation (9) is valid for the necessary and sufficient condition

$$\alpha_{kk} > \alpha_{km}. \tag{10}$$

(Note the difference between relation (10) and relation (4).) Also, from (8) and (3)

$$(p_k - p_{ck})\alpha_{ik} \leqq \sum_{m=0}^{n} (p_{cm} - p_m)\alpha_{im} \text{ for } i \neq k,\ m \neq k. \tag{11}$$

Therefore, in view of relation (6) it is necessary that

$$\alpha_{ik} \leqq \alpha_{im} \qquad \text{for } m \neq k,\ i \neq k. \tag{12}$$

Similarly, by interchanging the roles of $X_m$ and $X_k$, it can be shown that

$$\alpha_{ik} \geqq \alpha_{im} \qquad \text{for } k \neq m,\ i \neq m$$

whence

$$\alpha_{ik} = \alpha_{im} = \beta_i \quad \text{for } k \neq m,\ i \neq m,\ i \neq k. \tag{13}$$

The treatment, so far, has still allowed some variability in the scores ($\alpha_{ki}$). Next, let us suppose that the class of $N$ days is such that the true likelihoods of both $X_k$ and $X_0$ increased in proportion to their climatic frequencies. That is,

$$\frac{p_k}{p_{ck}} = \frac{p_0}{p_{c0}} > 1. \tag{14}$$

By the definition of skill, it is necessary, for $N$ large, that the forecasting of $X_k$ should net the forecaster the same true average score as the forecasting of $X_0$, both scores greater than 1.0. That is, using relations (13) and (5),

$$p_0\alpha_{00} = p_k\alpha_{kk} + (1 - p_k)\beta_k. \tag{15}$$

From (13), (5) and (3)

$$p_{c0}\alpha_{00} = 1. \tag{16}$$

From (13) and (3)

$$\alpha_{kk} = \frac{1 - \beta_k(1 - p_{ck})}{p_{ck}}. \tag{17}$$

From (14), (15), (16), (17) $\beta_k=0$; or

$$\alpha_{km} = 0 \qquad \text{for } m \neq k. \tag{18}$$

Returning to the $(n+1)$ relations (3), it follows that

$$\alpha_{ii} = \frac{1}{p_{ci}}. \tag{19}$$

Hence, *when the forecast requirement is to choose a single event from $(n+1)$ mutually exclusive events, the skill score for a perfect forecast is inversely proportional to the climatic frequency of the observed event;* the score for an incorrect forecast, however close to the observed event, is zero.

Professor Joseph G. Bryan, Massachusetts Institute of Technology, who collaborated in the preparation of [3], explained, in a personal interview, that the above result confirms the convictions that they had felt after their investigations of scores. Whenever they attempted to give partial credit for a nearly correct forecast, they compromised the system to the point where it was possible to "play" the system, usually by forecasting the most frequent event or by forecasting persistence.

It can be proved that the true average score for a forecast of any event $X_k$, whose $p_k > p_{ck}$, will be greater than unity. However, from (19) it follows that the true average score will be greatest for the event for which $(p_k - p_{ck})/p_{ck}$ is greatest. That is, the forecaster should forecast the event whose likelihood is increased most in proportion to its climatic frequency.

Ordinarily it is easier to make a prediction of the weather following one basic condition than it is to make the same prediction following another basic condition. For this reason persistence forecasts are oftentimes good forecasts; the forecaster is more likely to be correct

about clear skies six hours later if he observes clear skies at the time of his forecast. A table of scores, therefore, should be computed for each basic class of the weather (e.g. Table 5). If $X_k$ denotes a class of cloud heights, then $X_m'$ denotes the class of the cloud height at deadline time when the forecast is due. The climatic frequencies $p_{c0}, p_{c1}, \cdots, p_{cn}$ should be determined following each class of basic weather. If this is done, then, through relation (3), the true average score for "blind" forecasting will become unity (1.0), regardless of whether the "blind" forecast is a perpetual forecast of one event, or a forecast of persistence, or a forecast by pure chance.

In $s$ trials the forecaster's rating $F$ is the ratio of his accumulated score to the accumulated score for perfect forecasting:

$$F = \frac{\sum^{s} \alpha_{ki}}{\sum^{s} \alpha_{ii}}. \tag{20}$$

If it is felt that the rating for skill, $R$, should be zero for blind forecasting and unity for perfect forecasting, then the formula is

$$R = \frac{\sum^{s} \alpha_{ki} - s}{\sum^{s} \alpha_{ii} - s}. \tag{21}$$

Here, then, is a system of verification and scoring which requires little effort in procedure. However, it is too stark for a verification program involving competition among forecasters. If applied, it would yield average scores that would be too erratic, unless the competition were extended over a very long period of time. It seems better, for practical purposes, to compromise with rigor, to accept relations (3, 4, 5, 7), and (10) but to reject relation (8) in the sense that it will not be met rigorously. If this is done, then one can still be assured, by relation (7), that a skillful forecaster will have a true average score greater than unity, and, by equation (3), that the "blind" forecaster will have a true average score of unity. It is clear, therefore, that the forecaster will obtain a true average score greater than unity only if he has knowledge and experience with the weather sufficient to enable him to reckon with the increased or decreased likelihood of each future event.

The left-hand side of relation (8) may be greater than unity, but the scoring system will not suffer if

$$\sum_{m=0}^{n} p_m \alpha_{km} > \sum_{m=0}^{n} p_m \alpha_{im} \qquad \text{for } i \neq k. \tag{22}$$

Upon assigning adjusted scores, such as those in Table 6, this writer has not been able to "play" the system, that is, to extract a high average score out of Table 6 by forecasting a less-likely-than-usual event rather than an event of increased likelihood.

As an example, let us choose the flying weather classification at Randolph Field, Texas (Table 3). Let us consider only the winter months of December, January and February, the forecast to be made

TABLE 3. Flying Weather Classification. $H$ denotes ceiling height (ft), $V$ denotes visibility (miles)

| Symbol | Item | Description |
|---|---|---|
| $X_1$ | Airport Closed | $V<1$ or $0 \leqq H<500$ or both |
| $X_2$ | Instrument Flight Rules | $1 \leqq V<3$ and $H \geqq 500$ or $V \geqq 3$ and $500 \leqq H<1{,}000$ |
| $X_3$ | Contact with Low Clouds | $V \geqq 3$ and $1{,}000 \leqq H \leqq 2{,}000$ |
| $X_4$ | Contact; low clouds absent | $V \geqq 3$ and $H>2{,}000$ |

at deadline time 1230 CST, for the minimum conditions in the subsequent morning hours 0430 to 1030 CST. From 10 years of climatological data, the frequencies have been established as shown in Table 4. Solutions of equations (3, 18, and 19) yield the scores shown in Table 5 in which some of the scores are large while most of the scores are zero. Table 6, on the other hand, was prepared subject to the relations

TABLE 4. Frequency of the lowest flying weather classification between 0430 and 1030 CST in Dec., Jan., Feb., at Randolph Field, Texas

| Classification at 1230 CST | Frequency of classification between 0430 and 1030 CST | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1'$ | 60% | 14% | 6% | 20% |
| $X_2'$ | 46% | 18% | 20% | 16% |
| $X_3'$ | 33% | 20% | 20% | 27% |
| $X_4'$ | 18% | 10% | 9% | 63% |

TABLE 5. Ideal skill scores for each combination of forecast and verification of the flying weather classification at Randolph Field, Texas, Dec., Jan., or Feb. Forecast time is 1230 CST; verification period is between 0430 and 1030 CST

Antecedent Class: $X_1'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 1.7 | 0 | 0 | 0 |
| $X_2$ | 0 | 7.1 | 0 | 0 |
| $X_3$ | 0 | 0 | 16.7 | 0 |
| $X_4$ | 0 | 0 | 0 | 5.0 |

Antecedent Class: $X_2'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 2.2 | 0 | 0 | 0 |
| $X_2$ | 0 | 5.6 | 0 | 0 |
| $X_3$ | 0 | 0 | 5.0 | 0 |
| $X_4$ | 0 | 0 | 0 | 6.3 |

Antecedent Class: $X_3'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 3.0 | 0 | 0 | 0 |
| $X_2$ | 0 | 5.0 | 0 | 0 |
| $X_3$ | 0 | 0 | 5.0 | 0 |
| $X_4$ | 0 | 0 | 0 | 3.7 |

Antecedent Class: $X_4'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 5.6 | 0 | 0 | 0 |
| $X_2$ | 0 | 10.0 | 0 | 0 |
| $X_3$ | 0 | 0 | 11.1 | 0 |
| $X_4$ | 0 | 0 | 0 | 1.6 |

TABLE 6. Adjusted skill scores for each combination of forecast and verification of the flying weather classification at Randolph Field, Texas, Dec., Jan., or Feb. Forecast time is 1230 CST; verification period is between 0430 and 1030 CST

Antecedent class: $X_1'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 1.5 | 0.7 | 0.3 | 0 |
| $X_2$ | 1.0 | 1.9 | 0.9 | 0.5 |
| $X_3$ | 0.7 | 1.3 | 2.6 | 1.3 |
| $X_4$ | 0 | 1.0 | 1.9 | 3.8 |

Antecedent Class: $X_2'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 1.7 | 0.8 | 0.4 | 0 |
| $X_2$ | 0.9 | 1.8 | 0.9 | 0.5 |
| $X_3$ | 0.5 | 1.0 | 2.1 | 1.0 |
| $X_4$ | 0 | 0.8 | 1.6 | 3.3 |

Antecedent Class: $X_3'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 2.1 | 1.0 | 0.5 | 0 |
| $X_2$ | 0.9 | 1.9 | 0.9 | 0.5 |
| $X_3$ | 0.5 | 0.9 | 1.9 | 1.0 |
| $X_4$ | 0 | 0.6 | 1.2 | 2.4 |

Antecedent Class: $X_4'$

| Forecast | Verification | | | |
|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
| $X_1$ | 4.0 | 2.0 | 1.0 | 0 |
| $X_2$ | 1.3 | 2.6 | 1.3 | 0.6 |
| $X_3$ | 0.5 | 1.0 | 2.0 | 1.0 |
| $X_4$ | 0 | 0.5 | 0.7 | 1.4 |

(3, 4, 5) and (10); but the conditions (18) and (19) were replaced by the arbitrary relations

$$\alpha_{ki} = \frac{\alpha_{ii}}{2^{|k-i|}} \cdot \tag{23}$$

In the Air Force Geophysical Research Directorate, a trial project was chosen, the forecast of minimum flying weather classification at Randolph Field, Texas between the hours of 0430 CST and 1030 CST inclusive (Table 3). There were two daily forecasts which were due at deadlines 1230 CST of the day before and 0330 CST of the same day. For the first deadline, the table of scores was Table 6; for the second deadline a similar table was constructed.

TABLE 7a. Ratings for skill (R); daily forecasts between 1 Dec. 1949 and 28 Feb. 1950 of lowest flying weather classification at Randolph Field, Texas in the morning hours 0430–1030 CST

| Forecaster | Deadline 1230 CST (67 days) | Deadline 0330 CST (70 days) |
|---|---|---|
| A | 0.55 | 0.65 |
| B | 0.25 | 0.22 |
| C | 0.38 | 0.30 |

Note: No-skill score is zero; perfect score is 1.00. Standard deviation of each score is approximately $\pm 0.07$.

TABLE 7b. Forecaster B's ratings for skill and percentage of hits in the forecasting of lowest flying weather classification at Randolph Field, Texas during the 90 days of the winter months

| Season | | Deadline 1230 CST | Deadline 0330 CST |
|---|---|---|---|
| 1948–49 | Rating for skill | 0.16 | 0.11 |
| | Percentage of hits | 47% | 70% |
| 1949–50 | Rating for skill | 0.17 | 0.18 |
| | Percentage of hits | 48% | 62% |

Three forecasters, A, B, C, made forecasts on 67 to 70 days between 1 Dec. 1949 and 28 Feb. 1950 (Table 7a). Forecaster B, who had made his forecasts for all 90 days in the 1949–50 winter season, had made similar forecasts for the 90 days of the previous winter season (Table 7b).

The point raised by Tables 7a and 7b is that, aside from random variations, the three forecasters did not improve their skill in the 0330 CST forecast as compared with their skill in the 1230 CST forecast. Forecaster B, who had been making his forecasts for operational use, had a relatively high percentage of "hits" (Table 7b); moreover, his percentage of hits increased for the shorter-range forecasts. But, judged from the point of view of skill, his forecasts did not display more skill for the shorter-range forecasts than for the longer-range forecasts.

The above method of verification and scoring should prove satisfactory for the forecasting of certain weather elements such as temperature, pressure and amount of rainfall, which vary continuously from low values to high values. However, the method will prove unsatisfactory for the verification of an element such as "weather phenomenon." If he forecasts rain, the forecaster should not necessarily obtain credit when fog develops, wherefore the following requirement is now considered:

2a. *Requirement: One or more choices from* $(n+1)$ *mutually exclusive events; test of operational value.*

Let $\alpha_{k\ldots j}$ be the score for a correct forecast of the combination of events $X_k$, $X_{k+1}$, $\cdots$, $X_j$ treated as a single event. It is the job of the operations office to choose reasonable values for the $\alpha_{k\ldots j}$'s. If the weather is classed as "no weather," "fog" or "precipitation," the operations office might consider the forecast of fog as useless when rain verifies, and vice versa. A forecast of fog and precipitation as alternate possibilities might carry some weight, although less than the forecast of either event separately.

2b. *Requirement: One or more choices from* $(n+1)$ *mutually exclusive events; tests of skill.*

This requirement lends itself to a satisfactory system of scoring for skill. If $p_{c0}$, $p_{c1}$, $\cdots$, $p_{cn}$ are the climatic frequencies of each future event, for the given period of the year, for the given time of day, and for the given basic antecedent condition, $X_m'$, then, as in section 1b above, to make it no more profitable for the "blind" forecaster to predict one combination than another,

$$\alpha_{k\ldots j} \sum_{k}^{j} p_{cm} = 1. \tag{24}$$

Or, the score $\alpha_{k\ldots j}$ is inversely proportional to the climatic frequencies of the events $X_k$, $X_{k+1}$, $\cdots$, $X_j$ treated as a single event.

As an example, let us consider the weather phenomenon at Mitchell Field, L.I., classifying the weather as in Table 8. Let the forecasting be done in the month of July, deadline at 1430 EST, the forecast for 0230 EST of the following day. From 12 years of data, the frequency of the events at 0230 EST of the following day were found as in Table 9. The derived scores are shown in Table 10.

TABLE 8. Classification of Weather Phenomenon

| Symbol | Description of Class |
|---|---|
| $X_0$ | No weather |
| $X_1$ | Obstruction to vision other than fog |
| $X_2$ | Fog |
| $X_3$ | Precipitation |

TABLE 9. Frequency of each class of weather phenomenon at 0230 EST at Mitchell Field, L.I., July

| Class of weather phen. at 1430 CST of preceding day | Classification | | | |
|---|---|---|---|---|
| | $X_0$ | $X_1$ | $X_2$ | $X_3$ |
| $X_0'$ | 59.0% | 15.1% | 17.2% | 8.7% |
| $X_1'$ or $X_2'$ | 27.8% | 34.8% | 29.6% | 7.8% |
| $X_3'$ | 17.6% | 17.6% | 58.9% | 5.9% |

TABLE 10. Score for each combination of forecast at 1430 EST and verification of weather phenomenon at 0230 EST of following day at Mitchell Field, L.I., July

| Score | Antecedent Classification | | |
|---|---|---|---|
| | No weather $X_0'$ | Obst$^n$ to vision, or fog $X_1'$ or $X_2'$ | Precipitation or thunderstorm $X_3'$ |
| $\alpha_0$ | 1.7 | 3.6 | 5.7 |
| $\alpha_1$ | 6.6 | 2.9 | 5.7 |
| $\alpha_2$ | 5.8 | 3.4 | 1.7 |
| $\alpha_3$ | 11.5 | 12.8 | 17.0 |
| $\alpha_{01}$ | 1.4 | 1.6 | 2.8 |
| $\alpha_{12}$ | 3.1 | 1.6 | 1.3 |
| $\alpha_{23}$ | 3.9 | 2.8 | 1.5 |
| $\alpha_{012}$ | 1.1 | 1.1 | 1.1 |
| $\alpha_{123}$ | 2.4 | 1.4 | 1.2 |

The forecaster's rating, $F$, is the ratio of the accumulated score of the forecaster to the accumulated score for perfect forecasting:

$$F = \frac{\sum^{t} \alpha_{k \cdots j}}{\sum^{s} \alpha_i}$$

where $t$ is the number of "hits" in $s$ trials. If it is felt that the rating should be zero for "blind" forecasting and unity for perfect forecasting, then in $s$ trials, the rating for skill is

$$R = \frac{\sum^{t} \alpha_{k \cdots j} - s}{\sum^{s} \alpha_i - s}.$$

3*a*. *Requirement: Dichotomous grouping of the* $(n+1)$ *events; test of operational value.*

A requirement of this nature might become important at an airport where several types of aircraft are used, or where the pilots have varying flying qualifications. The forecaster must state whether the ceiling will be above the minimum for one airplane and pilot, above the minimum for another, and so on. We might suppose that $\alpha_k$ is the score for a correct forecast of ceiling equal to or below $X_k$, and $\beta_k$ the score for a correct forecast above $X_k$. It is the job of the operations office to choose reasonable values for the $\alpha_k$'s and $\beta_k$'s.

3*b*. *Requirement: Dichotomous grouping of the* $(n+1)$ *events; test of skill.*

Again, let us use the forecast of ceiling as an example. It is necessary that the forecaster make $n$ statements: first, as to whether the ceiling will be below or above $X_0$, secondly, whether below or above $X_1$, and so on. If $p_{c0}, p_{c1}, \cdots, p_{cn}$ are the climatic frequencies of $X_0$, $X_1, \cdots, X_n$ respectively, then it becomes equally profitable to the "blind" forecaster to predict ceilings above $X_k$ or equal to or below $X_k$ if

$$\alpha_k \sum_{m=0}^{k} p_{cm} = \beta_k \left(1 - \sum_{m=0}^{k} p_{cm}\right).$$

If

$$\alpha_k + \beta_k = 10,$$

then

$$\alpha_k = 10\left(1 - \sum_{m=0}^{k} p_{cm}\right) \quad \text{and} \quad \beta_k = 10\sum_{m=0}^{k} p_{cm}.$$

Such a system has the advantage that scores will range between zero and 10; but the forecaster must make $n$ statements on the weather.

## III. REMARKS

This paper, which grew out of a re-examination of previous programs of verification, was written in an effort to correct the previous weaknesses before another large-scale test of the abilities of forecasters is tried. One primary conclusion is that, to test their skill, forecasters should make special forecasts, independent of their operational duties, which forecasts might be inconsistent with the operational forecasts. There is, however, one feature common to the two forecasts: the forecaster's ability to analyze the antecedent weather into classes within which the probabilities of each subsequent event will differ substantially from its climatic frequency. It is this feature which should make a test of skill a good criterion of the usefulness of the forecaster's techniques in most operational forecasts.

Space will not permit a careful appraisal of the applicability of the above methods of scoring. I will simply say that tabulating machine methods could make a program of verification and scoring feasible with relatively few man-hours of labor. It would be possible to determine the relative merits of two or more forecasters, to determine how much easier it is to meet one requirement at one hour than it is to meet the same requirement at another hour, to determine whether it is more difficult to make forecasts in one season than in another, whether one forecaster has an advantage over another by being close to the station for which he is forecasting, and to determine whether it is more difficult to make a forecast under one prevailing condition than it is under another, such as continuous rain.

As a last remark, the principles and methods of this paper could easily apply to periodical forecasts of any kind. That is why this article appears in a statistical instead of a meteorological journal.

## REFERENCES

[1] Bleeker, W., "The Verification of Weather Forecasts," *Mededeelingen en Verhandelingen*, Serie B, Deel 1, no. 2 (1946), 23 pp.

[2] Muller, R. H., "Verification of Short Range Weather Forecasts" (A survey of the literature), *Bulletin of the American Meteorological Society*, Vol. 25 (1944), pp. 18–25, 47–53, 88–95.

[3] U. S. Army Air Forces, "Critique of Verification of Weather Forecasts," *Weather Division, USAAF*, Washington, D. C., 1944, 39 pp.
[4] Doolittle, M. H., "The Verification of Predictions," *American Meteorological Journal*, Vol. 2 (1884), pp. 327–29.
[5] Finley, J. P., "Tornado Predictions," *American Meteorological Journal*, Vol. 1, 1884, pp. 85–88.
[6] Köppen, W., "Einer rationelle Methode zur Prufung der Wetterprognosen," *Meteorologische Zeitschrift*, Vol. 1 (1884), pp. 397–404.
[7] Peirce, C. S., "The Numerical Measure of the Success of Predictions," *Science*, Vol. 4 (1884), pp. 453–54.
[8] Brier, G. W., "Verification of a Forecaster's Confidence and the Use of Probability Statements in Weather Forecasting," *U. S. Weather Bureau*, Dept. of Commerce, Washington, D. C., 1944, 10 pp.
[9] Brier, G. W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, Vol. 78 (1950), pp. 1–3.
[10] U. S. Air Force—Air Weather Service, "Critique of the Terminal Verification Program," 1949 (Typewritten; probable author, Capt. D. H. Russell).
[11] Clayton, H. H., "Verification of Weather Forecasts." *American Meteorological Journal*, Vol. 6 (1889), pp. 211–19.
[12] Heidke, P., "Berechnung des Erfolges und der Gute der Windstärkevorhersagen in Sturmwarnungsdienst," *Geografike Annaler*, Vol. 8 (1926), pp. 310–49.
[13] Gallé, P. H., "Étude Critique sur la methode de prevision du temps de Guilbert," *Mededeelingen en Verhandelingen*, Netherlands Meteorological Institute, Vol. 12 (1912).
[14] Grohmann, "Welchen Wert haben die Prüfungsergebnisse der Wetter vorhersagen öffentlicher Wetterdienstellen?" *Das Wetter*, Vol. 25 (1908), pp. 118–20.
[15] U. S. Army Air Forces, Weather Information Branch, "Short Range Forecast Verification Program," Report no. 602, Washington, D. C., 1943.
[16] Green, F. H. W., "Meteorological Forecasts for Special Purposes," *Nature*, Vol. 157 (1946), p. 556.
[17] Oxford University, "The Oxford University Dictionary," Oxford, 1933, p. 430.
[18] Gringorten, Irving I., "Forecasting by Statistical Inferences," *Journal of Meteorology*, Vol. 7 (1950), pp. 388–94.